SUPPLEMENTARY INFORMATION

1 Supplementary text

The two globally anti-correlated clusters seen in Figure 1D can be interpreted functionally by relating them to the conditions under which the genes are expressed. To this end, a statistical method extending principal component analysis, known as "singular value decomposition", can be applied, which reorders the genes and the conditions in a consistent way, according to their main axes of variation [1]. Specifically, the singular value decomposition of the transcription profile matrix $\bar{a}_{si}$ is of the form $\bar{a}_{si} = \sum \rho_k u_{sk} v_{ik}$, with $\rho_1 \geq \rho_2 \geq \cdots \geq 0$ the set of singular values. $\{u_s\}_{s=1...\#\text{conditions}}$ and $\{v_i\}_{i=1...\#\text{genes}}$ are, respectively, orthonormal basis of the gene space and of the condition space. The top singular vectors U_1 and V_1 have components $(U_1)_s = u_{s1}$ and $(V_1)_i = v_{i1}$, and define the main axes of variation in the gene space and condition space, respectively.

As a result, we obtain an ordered list of genes with the most anti-correlated genes at the two extremes, and an ordered list of conditions depending on whether they induce one or the other set of genes (Figure S1B-C). These lists indicate a simple interpretation of the two globally anti-correlated gene clusters in terms of phase of cell growth. Indeed, one gene cluster is preferentially expressed during exponential growth and the other during stationary phase (Figure S1D). This association of different growth rates with different overall patterns of gene expression is well recognized [2]. The preferential location of the anti-correlated genes on different halves of the genome is consistent with previous analyses [3].

References

[1] O Alter, P O Brown, and D Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA*, 97:10101–10106, 2000.

[2] O Shoval, H Sheftel, G Shinar, Y Hart, O Ramote, A Mayo, E Dekel, K Kavanagh, and U Alon. Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science*, 336:1157–1160, 2012.

[3] P Sobetzko, A Travers, and G Muskhelishvili. Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. *Proc. Natl. Acad. Sci. USA*, 109:E42–E50, 2012.

[4] T Vora, A K Hottes, and S Tavazoie. Protein occupancy landscape of a bacterial genome . *Molecular Cell*, 35:247–253, 2009.

[5] Charles J Dorman. H-NS, the genome sentinel. *Nature Reviews Microbiology*, 5(2):157–161, February 2007.

[6] I Junier, E Besray Unal, E Yus, V Llorens, and L Serrano. Insights into the mechanisms of basal coordination of transcription using a genome-reduced bacterium. *Cell Systems, in press*, 2016.

[7] Blanca Taboada, Ricardo Ciria, Cristian E Martinez-Guerrero, and Enrique Merino. ProOpDB: Prokaryotic Operon DataBase. *Nucleic Acids Research*, 40(Database issue):D627–31, 2012.

[8] Xuejiao Jiang, Patrick Sobetzko, William Nasser, Sylvie Reverchon, and Georgi Muskhelishvili. Chromosomal "stress-response" domains govern the spatiotemporal expression of the bacterial virulence program. *mBio*, 6(3):e00353–15, 2015.

[9] P Nicolas, U Mader, E Dervyn, T Rochat, A Leduc, N Pigeonneau, E Bidnenko, E Marchadier, M Hoebeke, S Aymerich, D Becher, P Bisicchia, E Botella, O Delumeau, G Doherty, E L Denham, M J Fogg, V Fromion, A Goelzer, A Hansen, E Härtig, C R Harwood, G Homuth, H Jarmer, M Jules, E Klipp, L Le Chat, F Lecointe, P Lewis, W Liebermeister, A March, R A T Mars, P Nannapaneni, D Noone, S Pohl, B Rinn, F Rugheimer, P K Sappa, F Samson, M Schaffer, B Schwikowski, L Steil, J Stülke, T Wiegert, K M Devine, A J Wilkinson, J M Van Dijl, M Hecker, U Völker, P Bessieres, and P Noirot. Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. *Science*, 335:1103–1106, 2012.

[10] J J Faith, M E Driscoll, V A Fusaro, E J Cosgrove, B Hayete, F S Juhn, S J Schneider, and T S Gardner. Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic acids research*, 36(Database):D866–D870, 2007.

2 Supplementary Figures

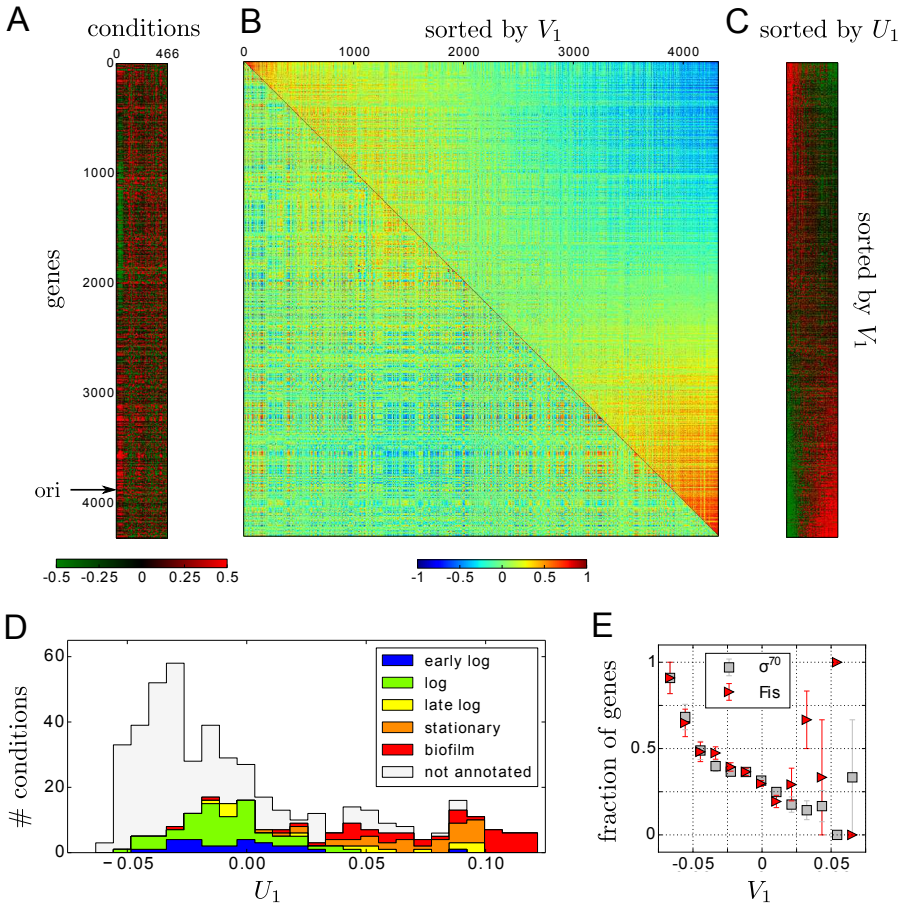

Figure S1: **A.** As in Figure 1A for *E. coli*, micro-array data reporting the expression levels of 4320 genes (rows) in 466 conditions (columns) with high expression in red and low expression in green. **B.** Applying a singular value decomposition to the micro-array data yields two principal components, V_1 along the genes and U_1 along the conditions. The co-expression matrix of Figure 1B is shown here with, above the diagonal, the genes sorted by V_1: this component classifies the genes according to their contribution to one of the two anti-correlated clusters visible in Figure 1D. **C.** Same expression data as in A, but with the conditions sorted by U_1 and the genes sorted by V_1, thus revealing the main pattern of variation. **D.** Distribution of the conditions along the principal component U_1, with different colors for the different phases of growth at which the measurements of transcriptional activity were made, showing that U_1 correlates with the growth rate. **E.** Fraction of genes controlled by σ^{70} (gray squares) and with a binding site for the NAP Fis (red triangles) as a function of V_1, showing that genes that are transcribed in growing phases (negative values of V_1) are more likely to be regulated by σ^{70} and bound by Fis.

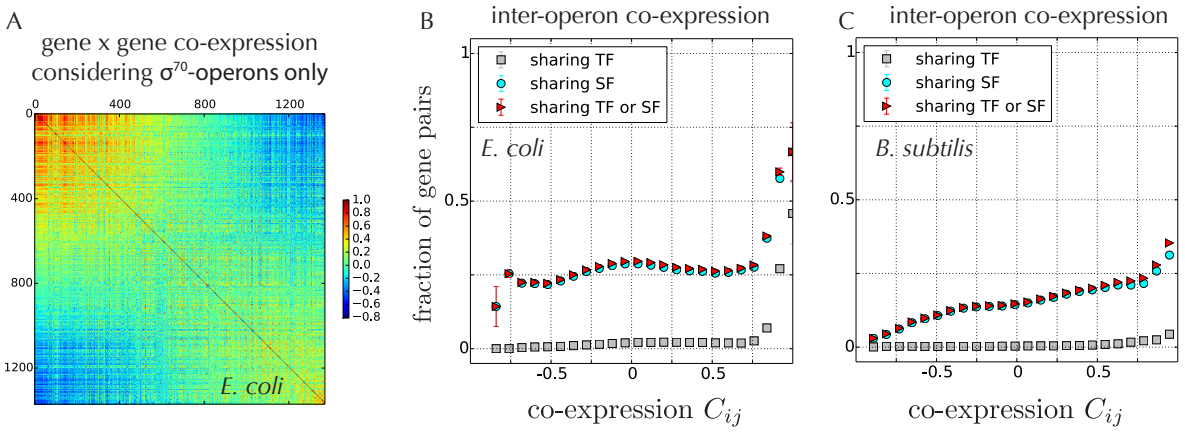

Figure S2: **A.** Transcriptional co-expression between the 1231 genes of *E. coli* having σ^{70} as unique SF. Genes are reordered along the first component V_1 from the SVD decomposition of the data as in Figure S1B. **B.** In *E. coli*, fraction of pairs of genes belonging to different operons that share a TF, a SF or one of the two, showing that, except at very high level co-expression ($C_{ij} > 0.85$), the majority ($\sim 75\%$) of correlated pairs of genes do not share a common TF or SF. **C.** Same analysis in *B. subtilis*.

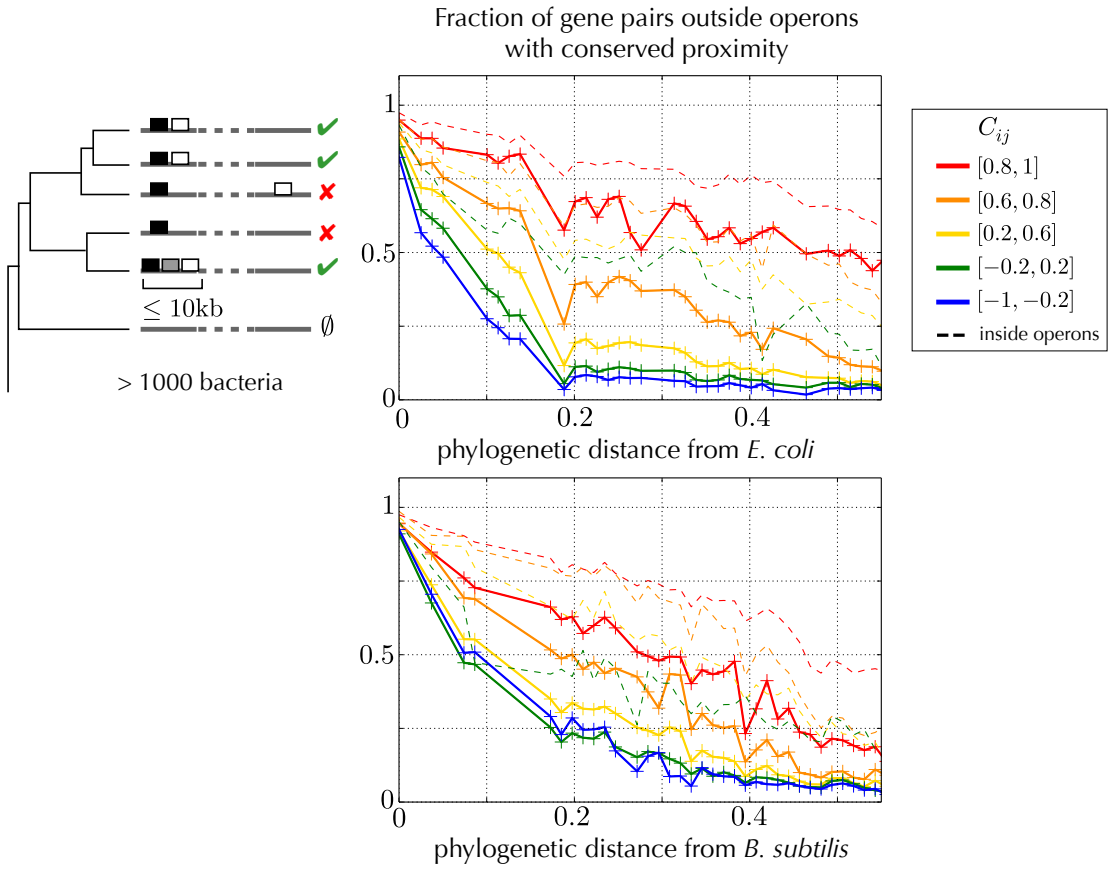

Figure S3: *Synteny as a proxy for high co-expression.* Taken two genes within 10 kb along the chromosome of a reference genome, what is the probability that they have orthologs within the same distance in the chromosome of another bacterium? We obtain an answer from a statistics over > 1000 bacterial genomes (left panel). This answer depends not only on the phylogenetic divergence between the query and reference genomes, but also very strongly on the level of co-expression of the two genes in the reference genome (plots): the more co-expressed are the two genes in *E. coli* (top) or in *B. subtilis* (bottom), the more likely they are to remain proximal in the chromosome of distant bacteria. The curves in the graph represent the fraction of pairs of genes within 10 kb in the reference genome (*E. coli* or *B. subtilis*) that are also within 10 kb in another genome as a function of the phylogenetic divergence between the two genomes (this divergence is measured by sequence divergence, see Materials and methods in main text). Different colors correspond to pairs of genes with different levels of co-expression in the reference genome: proximity between highly co-expressed pairs, in red, is thus much more conserved than between weakly co-expressed pairs, in yellow. The plain lines are based on pairs of genes that do not belong to the same operon, and the dotted lines on pairs of operonic genes: this shows that the relation between co-expression and synteny extends beyond operons.

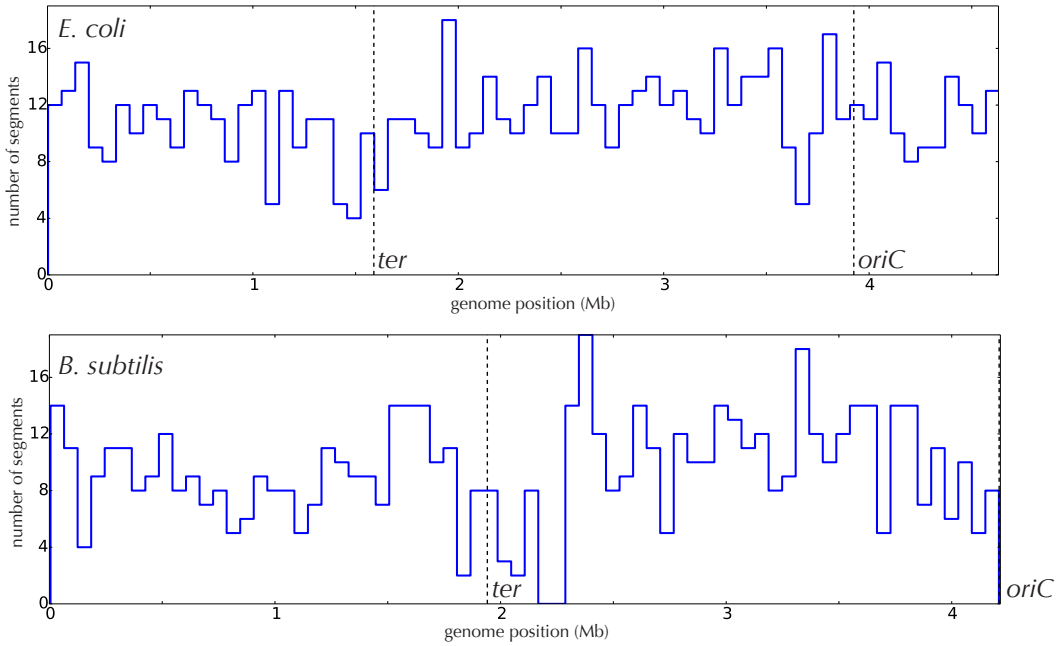

Figure S4: Genomic distribution of segments in *E. coli* (top) and in *B. subtilis* (bottom): the histograms of the location of the segments along the chromosome reveal a fairly uniform distribution (bin size of 65 kb). The vertical dashed lines indicate the origin (*oriC*) and terminus (*ter*) of replication. In *B. subtilis*, the depletion close to *ter* is mainly due to poor gene annotation in this region.

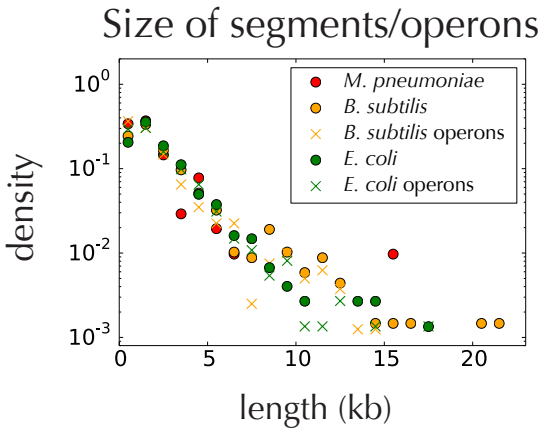

Figure S5: Size distributions of synteny segments (solid circles) in three phylogenetically distant bacteria and of polycistronic operons in *E. coli* and in *B. subtilis* (crosses), showing a similar exponential decrease up to $\sim$ 10 kb.

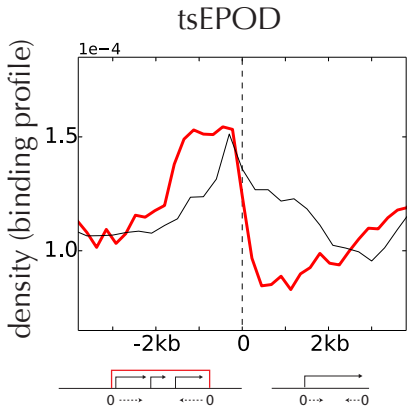

Figure S6: Binding profile of tsEPODs [4] with respect to synteny segments (red plain line) and operons (black), showing, as in the case of H-NS (Figure 2D in main text), a strikingly high density of tsEPODs at the external boundaries of segments together with a depletion inside segments. In agreement with their role in transcription silencing [5], we also observe an enrichment around the promoter region, and over the first gene for operons not at the border.

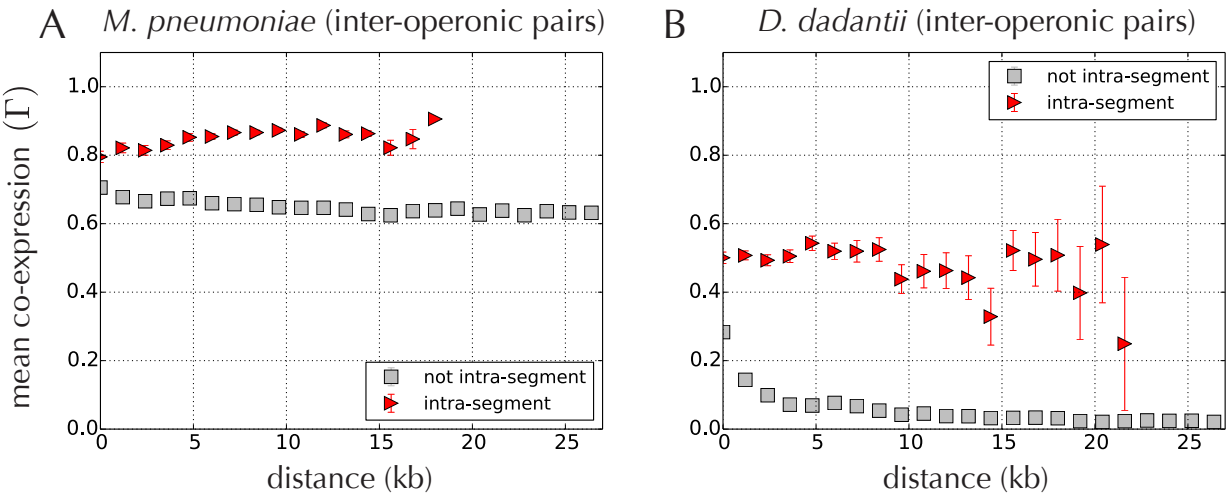

Figure S7: Co-expression analysis for two additional bacteria: **A.** *Mycoplasma pneumoniae* (classified as closed to Gram-positive) and **B.** *Dickeya dadantii* (formerly *Erwinia chrysanthemi*, Gram-negative). These two bacterial strains have very different genome lengths (they contain respectively ca. 650 and 4500 protein coding genes) and lifestyles (*M. pneumoniae* is a human parasit living in the respiratory tract, *D. dadantii* is a plant pathogen); they are also phylogenetically distant from both *E. coli* and *B. subtilis* (analyzed in Figure 4). *M. pneumoniae* is known to have a tiny repertoire of TFs and a single SF, while the regulatory network of *D. dadantii* is mostly unknown (as for most bacteria). The graphs compare co-expression inside synteny segments (red triangles) to co-expression outside segments (gray squares). In any case, only genes belonging to different operons are considered (operon map from [6] for *M. pneumoniae* and from the ProOpDB database [7] for *D. dadantii*). Co-expression levels are computed from rRNA normalized RNA-seq data obtained in 151 different conditions for *M. pneumoniae* [6] and from rRNA normalized micro-array data obtained in 32 different conditions for *D. dadantii* [8]. Although global levels of co-expression differ between strains (see [6] for a detailed analysis of co-expression properties in *M. pneumoniae*), a systematic enhancement of co-expression is observed inside synteny segments, which is nearly independent of the distance separating the genes.

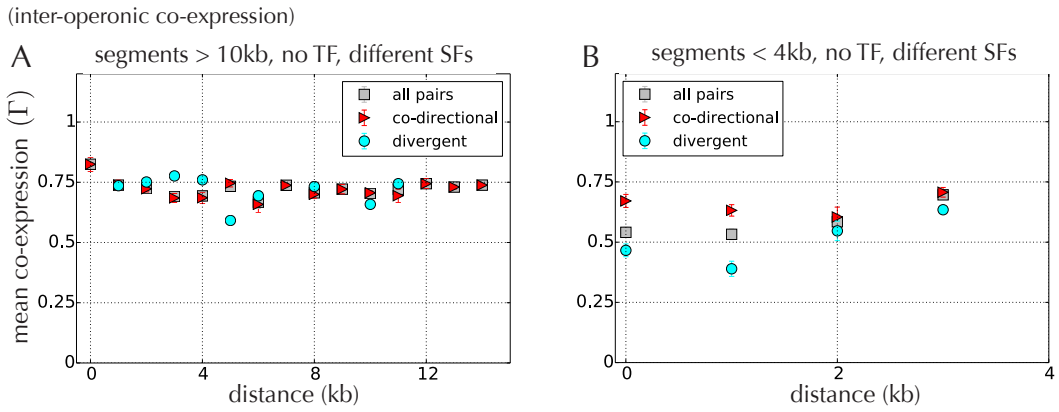

Figure S8: **A.** The red triangles correspond to those of Figure 4B (*E. coli*), and the gray squares and cyan points show that restricting to co-directional or divergent pairs has little incidence. **B.** Similar to A, but considering the smallest segments (< 4 kb) instead of the largest ones (> 10 kb): the overall level of correlation is lower for shorter segments.

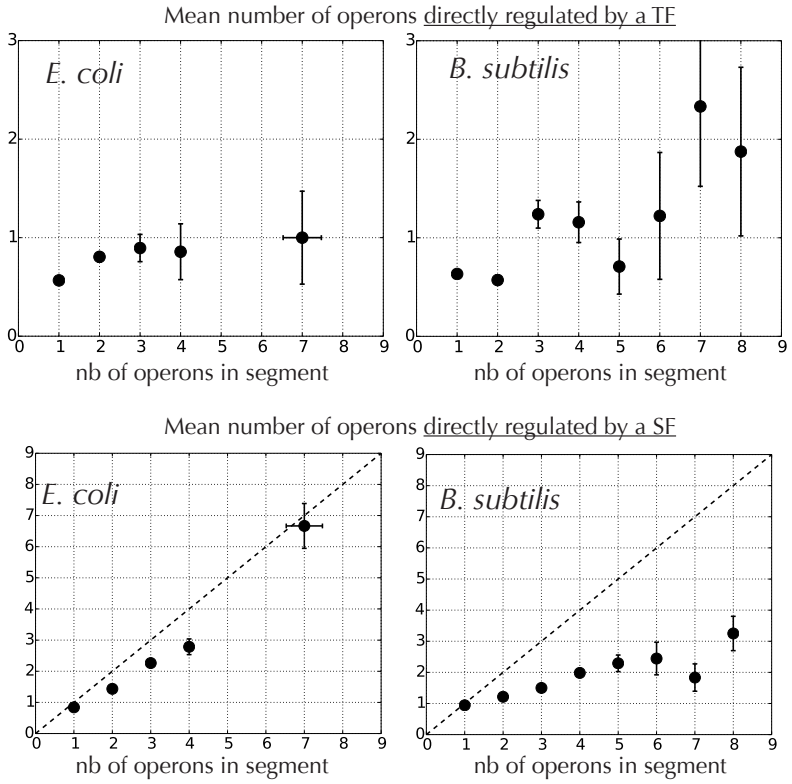

Figure S9: Average number of operons controlled by at least one TF (upper panels) or by at least one SF (lower panels) as a function of the number of operons in the segment. Results show that both in *E. coli* (left panels) and in *B. subtilis* (right panels) there is roughly a constant number (close to 1) of operons directly regulated by a TF. In contrast, most operons are directly regulated by a SF in *E. coli* (left lower panel). In *B. subtilis*, not all operons of the segment are regulated by a SF, but at least one. The dashed lines in the lower panels indicate the bisectors $y = x$.

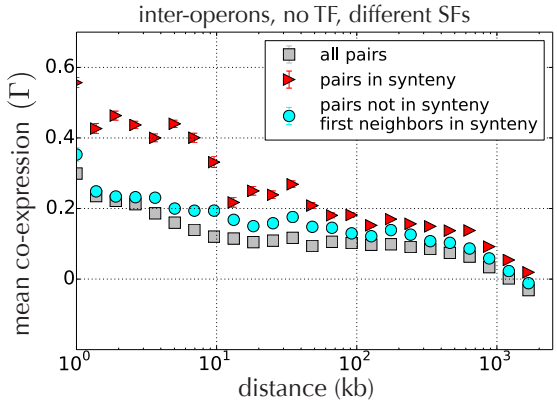

Figure S10: Co-expression between *E. coli* genes in different operons that are not regulated by any TF and that do not share the same SF (gray squares). Pairs in synteny, independently of whether they are proximal in the chromosome of *E. coli*, are on average more co-expressed than those not in synteny (red triangles). The phenomenon appears to be specific since replacing the first gene in these pairs by its nearest neighbor not in synteny (while keeping the second gene) systematically decreases the mean level of co-expression at all distances.

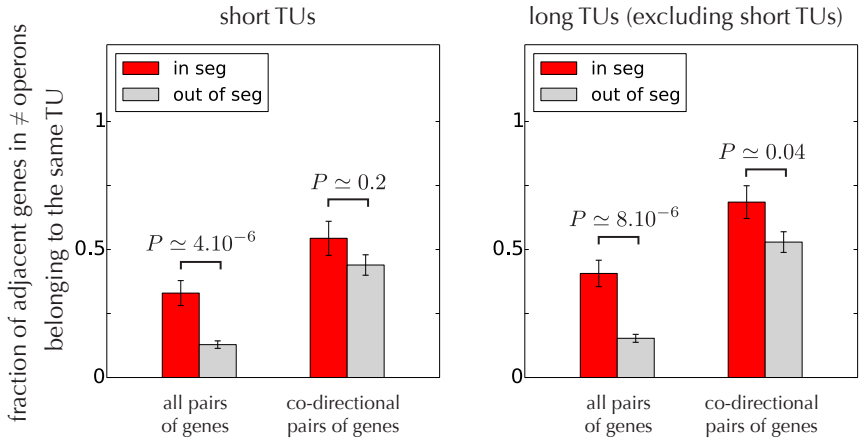

Figure S11: Fraction of adjacent genes that belong to a same transcriptional unit (TU) as identified in *B. subtilis* [9] (additional details Figure 6B). Two types of TUs are considered as proposed in [9]: "short TUs" (left panel), which are minimal TUs found in most conditions, and "long TUs" (right panel), which are maximal TUs found in at least one condition. The fraction is computed for genes inside synteny segments (red bars) and for genes outside synteny segments (gray bars). In each panel, the two bars on the left are based on all pairs of genes in different operons and those on the right on pairs of co-directional genes in different operons.

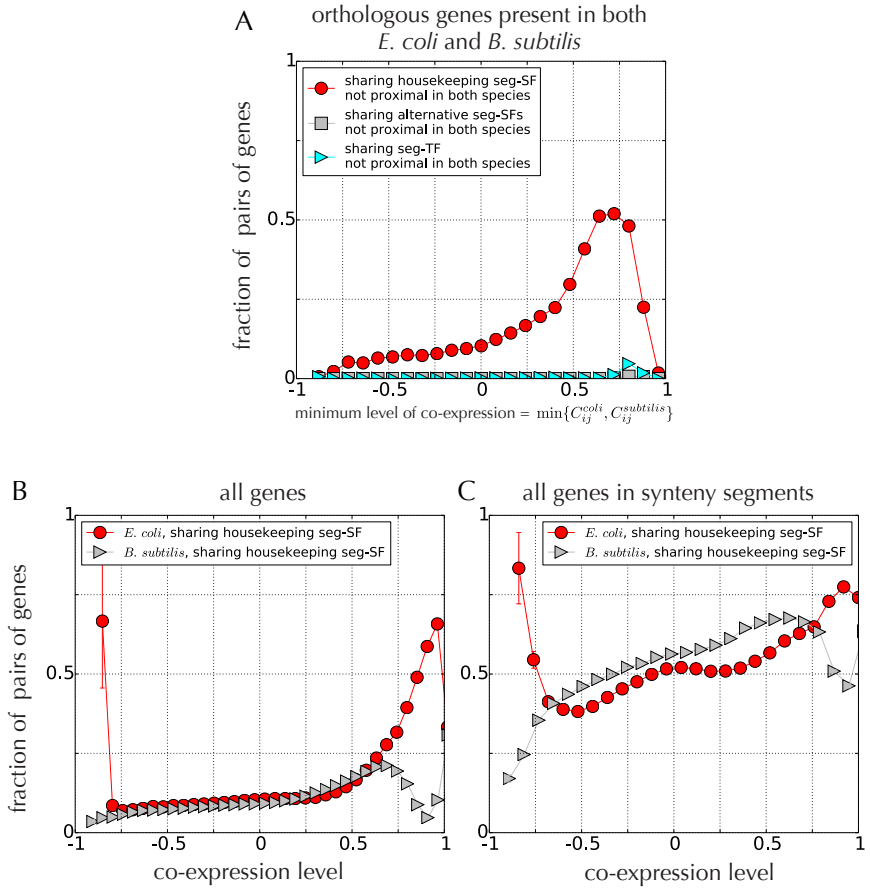

Figure S12: **A.** Extension of the results of Figure 6C, showing that conserved high co-expression is mostly due to a seg-regulation by housekeeping SFs (σ^{70} in *E. coli* and SigA in *B. subtilis*). **B.** Contribution of the seg-regulation by housekeeping SFs in each organism. **C.** Same as in B but considering only genes that belong to synteny segments, showing a strong relationship in both bacteria between gene co-expression and seg-regulation by a housekeeping SF. In B and C, the unexpected drop at high co-expression level for *B. subtilis* may either come from a too partial annotation of SF binding sites, or from the imperfect match between our synteny segments and the actual relevant co-expression unit of *B. subtilis*.

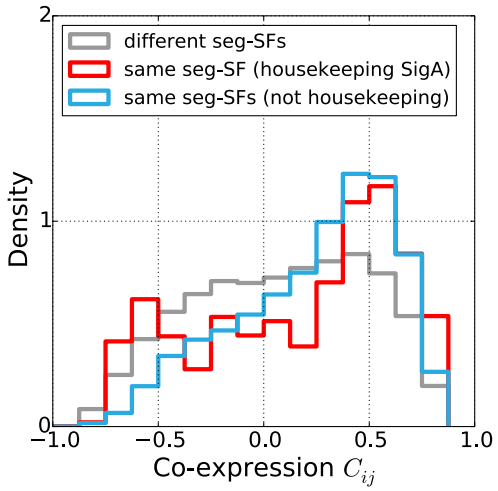

Figure S13: Distribution in *B. subtilis* of the co-expression C_{ij} between pairs of genes that are not directly regulated by a TF or a SF and that belong to different synteny segments. Gray distribution: pairs in segments with different sets of SFs. Red distribution: pairs in segments that have one single seg-SF, the housekeeping SigA. Cyan distribution: pairs in segments that have exactly the same seg-SFs, excluding SigA.

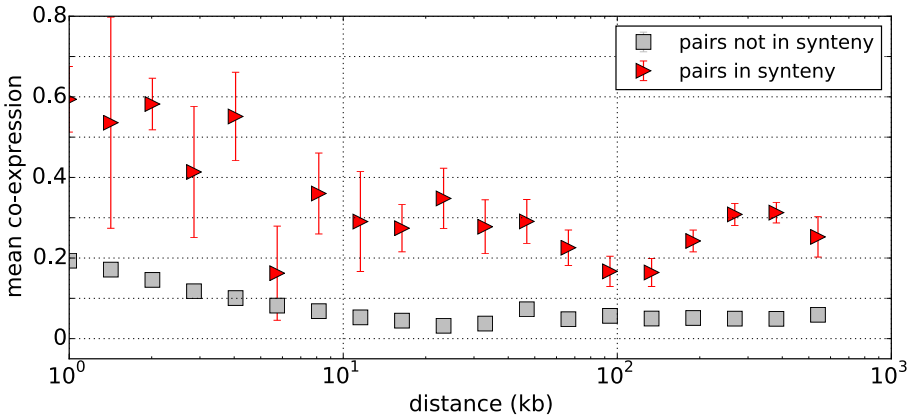

Figure S14: Co-expression for pairs of genes in synteny (red triangles) or not (gray squares) in *S. cerevisiae*. Synteny is defined as in the main text, from a dataset of bacterial genomes that does not include any yeast genome. Co-expression is computed from micro-array data retrieved from the M^{3D} database [10]. Pairs of genes in synteny in bacteria are in average more co-expressed in *S. cerevisiae* that pairs than are not in synteny in bacteria.

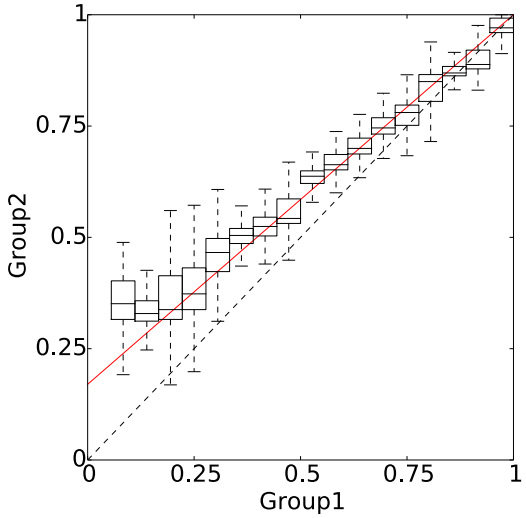

Figure S15: Robustness of the calculation of evolutionary distances. We compare two evolutionary distances that were computed using two different groups of 5 genes that reflect phylogenetic distances between bacterial strains (Materials and methods in main text) and with no gene in common. One can observe a linear relationship (in red) for almost the full range of similarities, except at very low similarities. All genome pairs formed from the 1445 genomes of our dataset are reported. The dashed black line indicates the bisector $y = x$.

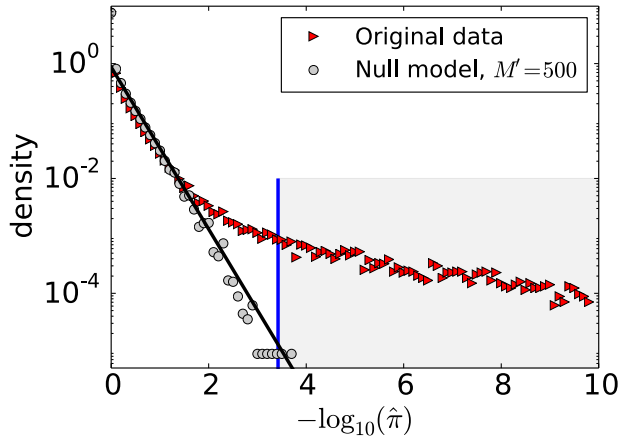

Figure S16: Probability density of $-\log(\hat{\pi})$ for the empirical data (red triangles) obtained for an effective number of genomes $M' = 500$. For small enough values of $-\log(\hat{\pi})$, the density decays exponentially with $-\log(\hat{\pi})$ (black line). The deviation from an exponential at large values (gray area) indicates the conservation of co-localization. For the null model (gray points), for which we consider the same effective number of genomes but where gene positions are randomized, the exponential decay extends to larger values of $-\log(\hat{\pi})$. Here, we consider a false discovery rate FDR = 0.005, leading to a threshold $\pi^* \simeq 4.10^{-4}$ (vertical blue line) – see Materials and methods in main text.